A PRACTICAL TECHNIQUE FOR ESTIMATING GENERAL REGRESSION SURFACES

by
Donald F. Specht

6-79-68-6 June 1968

Electronic Sciences Laboratory
Lockheed Palo Alto Research Laboratory
LOCKHEED MISSILES & SPACE COMPANY
A Group Division of Lockheed Aircraft Corporation
Palo Alto, California

ACKNOWLEDGMENTS

The author gratefully acknowledges the contributions of Professor H. Chernoff of Stanford University (Department of Statistics) and Dr. L. T. Stewart (Lockheed Research Laboratory). Their review and constructive criticism of the manuscript was most helpful. Technical discussions with Dr. N. Kusnezov and other colleagues at Lockheed have also been valuable.

This work was performed under the Lockheed Independent Research Program. The support and encouragement of Dr. C. E. Dur an ar greatly appreciated.

CONTENTS

Section 1
INTRODUCTION AND SUMMARY

Let $\underset{\sim}{X}$ be a p component random vector variable with transpose $\underset{\sim}{X}' \equiv [X_1, \ldots X_j, \ldots X_p]$ and Y be a one-dimensional random variable with a joint continuous distribution of density $f(\underset{\sim}{x}, y)$. The regression of Y given $\underset{\sim}{X} = \underset{\sim}{x}$ is

$$E[Y|\underset{\sim}{X} = \underset{\sim}{x}] = \int_{-\infty}^{\infty} yf(\underset{\sim}{x}, y)\, dy \div \int_{-\infty}^{\infty} f(\underset{\sim}{x}, y)\, dy \tag{1.1}$$

Using the consistent nonparametric estimators of the densities described later in the paper, the regression can be estimated by the estimated regression

$$\hat{Y}(\underset{\sim}{x}) \cong \frac{\sum_{i=1}^{n} Y_i \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{p}(X_{ji} - x_j)^2\right]}{\sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{p}(X_{ji} - x_j)^2\right]} \tag{1.2}$$

where

n = number of observations of $\underset{\sim}{X}$ and Y used in estimating the densities

$\underset{\sim}{X}_i$ and Y_i = the ith observations of $\underset{\sim}{X}$ and Y, respectively

$\underset{\sim}{X}_i' \equiv [X_{1i}, \ldots, X_{ji}, \ldots, X_{pi}]$

σ = a smoothing parameter

It is further shown that the nonlinear regression equation (1.2) can be approximated to any desired accuracy by a polynomial-ratio regression estimate

$$Y^*(\underset{\sim}{x}) = \frac{Q(\underset{\sim}{x})}{P(\underset{\sim}{x})} \tag{1.3}$$

where the coefficients of the polynomials are computed as a function of the observed sample. The advantage of the form Eq. (1.3) is that the observations are used only in the computation of the coefficients. Subsequent evaluation of $\hat{Y}(\underset{\sim}{x})$ for a given vector $\underset{\sim}{x}$ is usually much faster using Eq. (1.3) rather than Eq. (1.2).

This same advantage is, of course, shared by classical polynomial regression equations, but the technique described has the following advantages compared with classical polynomial regression techniques utilizing a single polynomial:

- It provides a simple method of determining the coefficients. The calculation for the coefficient of a particular term amounts to little more than averaging the corresponding product of variables over the set of observations available.
- The computational and storage requirements increase only linearly with the number of coefficients used.
- The shape of the regression surface can be made as complex as necessary to closely approximate Eq. (1.1), or as simple as desired, by proper choice of the smoothing parameter σ. In spite of this flexibility, $Y^*(\underset{\sim}{x})$ estimated from Eq. (1.3) is bounded by the minimum and maximum of the observations Y_i when $Q(\underset{\sim}{x})$ and $P(\underset{\sim}{x})$ are truncated at an even order.
- Because of the smoothing properties inherent in the density estimator, the number of coefficients used in the polynomials can approach or even exceed the number of observations in the sample with no danger of the regression surface overfitting the data when σ is suitably chosen.

Since the derivation of Eq. (1.3) from Eq. (1.2) is not dependent on $\underset{\sim}{X}$ being random, the computational advantages of the polynomial form can be utilized also when values of $\underset{\sim}{X}$ are specified in the design of an experiment and Y alone is a random variable. This property applies, of course, to ordinary polynomial regression as well.

Section 2
GENERAL REGRESSION

Let $\underset{\sim}{X}$ be a p component random vector variable with transpose $\underset{\sim}{X}' \equiv [X_1, \ldots X_j, \ldots X_p]$, and let Y be a random variable with a joint continuous distribution of density $f(\underset{\sim}{x}, y)$. The conditional mean of Y given $\underset{\sim}{X} = \underset{\sim}{x}$,

$$E[Y|\underset{\sim}{X} = \underset{\sim}{x}] = \int_{-\infty}^{\infty} yf(\underset{\sim}{x}, y)\, dy \div \int_{-\infty}^{\infty} f(\underset{\sim}{x}, y)\, dy \qquad (2.1)$$

is also called the regression of Y on $\underset{\sim}{X}$ and is determined by the density f.

When the density $f(\underset{\sim}{x}, y)$ is not known, it must usually be estimated from a sample of observations of $\underset{\sim}{X}$ and Y. We estimate the regression by taking the regression of a nonparametric estimate of $f(\underset{\sim}{x}, y)$. The class of consistent estimators proposed by Parzen (Ref. 1) and shown to be applicable to the multidimensional case by Cacoullos (Ref. 2) are suitable for this purpose. For reasons expressed in Refs. 3 and 4, the particular estimator

$$\hat{f}(\underset{\sim}{x}) = \frac{1}{\sigma^p (2\pi)^{p/2}} \frac{1}{n} \sum_{i=1}^{n} \exp\left[-\frac{(\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x})}{2\sigma^2}\right] \qquad (2.2)$$

seems to be a good choice for estimating a probability density function f if it can reasonably be assumed that the underlying density is continuous and that its first partial derivatives evaluated at any $\underset{\sim}{x}$ are small.

Letting $(\underset{\sim}{X}^*)' \equiv [X_1, \ldots, X_p, Y]$ and $(\underset{\sim}{x}^*)' \equiv [x_1, \ldots, x_p, y]$, the application of Eq. (2.2) yields the estimated regression

$$\hat{Y}(\underset{\sim}{x}) = \int_{-\infty}^{\infty} y\hat{f}(\underset{\sim}{x}^*)\, dy \div \int_{-\infty}^{\infty} \hat{f}(\underset{\sim}{x}^*)\, dy$$

$$= \frac{\int_{-\infty}^{\infty} y \sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(\underset{\sim}{X}_i^* - \underset{\sim}{x}^*)'(\underset{\sim}{X}_i^* - \underset{\sim}{x}^*)\right] dy}{\int_{-\infty}^{\infty} \sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(\underset{\sim}{X}_i^* - \underset{\sim}{x}^*)'(\underset{\sim}{X}_i^* - \underset{\sim}{x}^*)\right] dy}$$

$$= \frac{\sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x})\right] \int_{-\infty}^{\infty} y \exp\left[-\frac{(Y_i - y)^2}{2\sigma^2}\right] dy}{\sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x})\right] \int_{-\infty}^{\infty} \exp\left[-\frac{(Y_i - y)^2}{2\sigma^2}\right] dy}$$

Letting

$$A_i^2 = (\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x}) = \sum_{j=1}^{p} (X_{ji} - x_j)^2 \qquad (2.3)$$

and performing the indicated integrations,

$$\hat{Y}(\underset{\sim}{x}) = \frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{A_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{A_i^2}{2\sigma^2}\right)} \qquad (2.4)$$

Note that because the particular estimator Eq. (2.2) is readily decomposed into $\underset{\sim}{X}$ and Y factors, the integrations were accomplished analytically. The resulting regression equation (2.7) which involves summations over the observations is simply applicable to problems involving numerical data.

Although $\hat{Y}(\underset{\sim}{x})$ will be expressed in terms of series expansions to introduce a computationally more efficient approximation in the next section, the main properties of $\hat{Y}(\underset{\sim}{x})$ are evident in the present form.

Parzen (Ref. 1) and Cacoullos (Ref. 2) have shown that the density estimator $\hat{f}(\underset{\sim}{x})$ [Eq. (2.2)] used in estimating Eq. (2.1) by Eq. (2.4) is a consistent estimator [asymptotically converges to the underlying probability density function $f(x)$] at all points $\underset{\sim}{x}$ at which the density function is continuous, providing that $\sigma = \sigma(n)$ is chosen as a function of n such that

$$\lim_{n\to\infty} \sigma(n) = 0$$

and

$$\lim_{n\to\infty} n\sigma(n) = \infty$$

The estimate $\hat{Y}(\underset{\sim}{x})$ can be visualized as a weighted average of all of the observed values Y_i where each observed value is weighted exponentially according to its Euclidean distance from $\underset{\sim}{x}$. When the smoothing parameter σ is made large, the estimated density is forced to be smooth and in the limit becomes multivariate Gaussian with covariance $\sigma^2 I$. On the other hand, small σ allows the estimated density to assume non-Gaussian shapes but with the hazard that wild points may have too great an effect on the estimate. As $\sigma \to \infty$, $\hat{Y}(\underset{\sim}{x})$ assumes the value of the sample mean of the observed Y_i, and as $\sigma \to 0$, $\hat{Y}(\underset{\sim}{x})$ assumes the value of the Y_i associated with the observation closest to $\underset{\sim}{x}$.* (This case is treated in more detail in Ref. 5.) For

*Consider two observations $(\underset{\sim}{X}_1, Y_1)$ and $(\underset{\sim}{X}_2, Y_2)$ such that $A_2^2 = A_1^2 + \epsilon$, $\epsilon > 0$ for some value of x. Then from Eq. (2.4),

$$\hat{Y} = \frac{\exp\left(-\frac{A_1^2}{2\sigma^2}\right)\left[Y_1 + Y_2 \exp\left(-\frac{\epsilon}{2\sigma^2}\right)\right]}{\exp\left(-\frac{A_1^2}{2\sigma^2}\right)\left[1 + \exp\left(-\frac{\epsilon}{2\sigma^2}\right)\right]} \quad ; \quad \lim_{\sigma\to 0} \hat{Y} = Y_1$$

intermediate values of σ, all values of Y_i are taken into account, but those corresponding to points nearer to $\underset{\sim}{x}$ are given heavier weight.

When the underlying parent distribution is not known, it is not possible to compute an optimum σ for a given number of observations n. It is therefore necessary to find σ on an empirical basis. This can be done quite easily when the density estimate is being used in a regression equation because there is a natural criterion which can be used for evaluating each value of σ; namely, the correlation between Y_i and the estimate $\hat{Y}(\underset{\sim}{X}_i)$ for each of the observed samples. One precaution is necessary, however. For this purpose $\hat{Y}(\underset{\sim}{X}_i)$ must be modified to

$$\tilde{Y}(\underset{\sim}{X}_i) = \sum_{j \neq i} Y_j \exp\left(-A_j^2/2\sigma^2\right) \div \sum_{j \neq i} \exp\left(-A_j^2/2\sigma^2\right)$$

so that each $\hat{Y}(\underset{\sim}{X}_i)$ is based on inference from all the observations except the actual observed value at $\underset{\sim}{X}_i$. This procedure is used to avoid an artificial maximum correlation as $\sigma \rightarrow 0$ which results when the estimated density is allowed to fit the observed data points. (Overfitting of the data is also present in the least-squares estimation of linear regression surfaces, but is not as severe because the linear regression equation has only $p + 1$ degrees of freedom. If $n \gg p$, the phenomenon of overfitting can be and is commonly ignored in least-squares regression.)

Section 3
A POLYNOMIAL EQUIVALENT

3.1 DERIVATION

In Ref. 4 it was shown that since the density estimator $\hat{f}(\underset{\sim}{x})$ can be written

$$\hat{f}(\underset{\sim}{x}) = (2\pi\sigma^2)^{-p/2} \exp\left[-(\underset{\sim}{x}'\underset{\sim}{x})/2\sigma^2\right] \frac{1}{n} \sum_{i=1}^{n} \exp\left[\underset{\sim}{x}'\underset{\sim}{X}_i/\sigma^2\right] \exp\left[-\underset{\sim}{X}_i'\underset{\sim}{X}_i/2\sigma^2\right]$$

it can be replaced by a polynomial approximation based on a Taylor's series expansion of $\exp\left[\underset{\sim}{x}'\underset{\sim}{X}_i/\sigma^2\right]$. In many circumstances, this approximation requires substantially less computation than $\hat{f}(\underset{\sim}{x})$ [Eq. (2.2)]. The polynomial version of the estimator has the form

$$f_\ell^*(\underset{\sim}{x}) = (2\pi\sigma^2)^{-p/2} \exp\left[-(\underset{\sim}{x}'\underset{\sim}{x})/2\sigma^2\right] P_\ell(\underset{\sim}{x}) \tag{3.1}$$

where

$$P_\ell(\underset{\sim}{x}) = \sum_{J \le \ell} a_{j_1 \ldots j_p} x_1^{j_1} x_2^{j_2} \cdots x_p^{j_p} \quad , \quad j_i \ge 0$$

$$J = j_1 + j_2 + \ldots + j_p \tag{3.2}$$

The coefficients $a_{j_1 \ldots j_p}$ are computed from the observations $\underset{\sim}{X}_i$ using

$$a_{j_1 \ldots j_p} = \left[n\sigma^{2J} j_1!\, j_2! \cdots j_p!\right]^{-1} \sum_{i=1}^{n} \left[X_{1i}^{j_1} \cdots X_{pi}^{j_p} \exp\left(-\underset{\sim}{X}_i'\underset{\sim}{X}_i/2\sigma^2\right)\right] \tag{3.3}$$

and the sum is over all j for which $J \leq \ell$. Note that each coefficient $a_{j_1 \ldots j_p}$ involves the ith observation $\underset{\sim}{X}_i$ only in one of a sum of n terms.

Although the generality of the notation used makes the equations look formidable, consider the coefficient of a specific term in Eq. (3.2) such as the coefficient $a_{110\ldots0}$ of $x_1 x_2$. Then

$$a_{110\ldots0} = \frac{1}{n\sigma^4} \sum_{i=1}^{n} X_{1i} X_{2i} \exp\left(-\underset{\sim}{X}_i' \underset{\sim}{X}_i / 2\sigma^2\right)$$

In words, this equation says to take the average of the products of the cross products $X_{1i}X_{2i}$ and a "normalizing factor" $\exp\left(-\underset{\sim}{X}_i' \underset{\sim}{X}_i / 2\sigma^2\right)$; then to multiply this average by a "premultiplying constant" $1/n\sigma^4$. Each term has its own premultiplying constant. Note, however, that all terms for an observation have the same normalizing factor. The normalizing factor, therefore, need be calculated only once for each observation, regardless of the number of coefficients used in the polynomial $P_\ell(\underset{\sim}{x})$. Considering this circumstance, and also the fact that the premultiplying constant is not data-dependent, the algorithm implied by Eq. (3.3) amounts to little more computation than simply making each coefficient equal to the mean of the corresponding cross product over the observation set used for establishing the coefficients.

Note that the regression equation (2.4) can be written

$$\hat{Y}(\underset{\sim}{x}) = \frac{n^{-1}(2\pi\sigma^2)^{-p/2} \sum_{i=1}^{n} Y_i \exp\left[-(\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x})/2\sigma^2\right]}{n^{-1}(2\pi\sigma^2)^{-p/2} \sum_{i=1}^{n} \exp\left[-(\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x})/2\sigma^2\right]}$$

$$= \frac{n^{-1}(2\pi\sigma^2)^{-p/2} \sum_{i=1}^{n} Y_i \exp\left[-(\underset{\sim}{X}_i - \underset{\sim}{x})'(\underset{\sim}{X}_i - \underset{\sim}{x})/2\sigma^2\right]}{\hat{f}(\underset{\sim}{x})} \qquad (3.4)$$

A polynomial approximation for the denominator of Eq. (3.4) has just been given; a similar polynomial approximation can be derived for the numerator of this expression. When these are both used in Eq. (3.4), we approximate $\hat{Y}(\underset{\sim}{x})$ by the polynomial-ratio regression estimate

$$Y^*_\ell(\underset{\sim}{x}) = \frac{Q_\ell(\underset{\sim}{x})}{P_\ell(\underset{\sim}{x})} \tag{3.5}$$

where $Q_\ell(\underset{\sim}{x})$ has identically the same form as $P_\ell(\underset{\sim}{x})$ except that the coefficients are computed by

$$b_{j_1 \ldots j_p} = \left[n\sigma^{2J} j_1! \ldots j_p!\right]^{-1} \sum_{i=1}^{n} \left[Y_i X_{1i}^{j_1} \ldots X_{pi}^{j_p} \exp\left(-\underset{\sim}{X}_i'\underset{\sim}{X}_i/2\sigma^2\right)\right] \tag{3.6}$$

3.2 NORMALIZATION OF INPUT

Since the accuracy of a finite term Taylor's series expansion of $\exp(\underset{\sim}{x}'\underset{\sim}{X}/\sigma^2)$ depends on the magnitude of $\underset{\sim}{x}'\underset{\sim}{X}/\sigma^2$, it is often desirable to make a transformation (translation) from raw measurement vectors (e.g., $\underset{\sim}{Z}_i$ for the ith observation) to obtain a set of vectors $\underset{\sim}{X}_i$ for which the Taylor approximation is satisfactory. Similarly, to minimize distortion of the estimated density relative to the parent density and simultaneously to minimize error due to the Taylor's series expansion, it is desirable to "sphericalize" the data in some way. The simplest procedure consists of normalizing each variate to have a variance of unity. If the variables are highly interdependent, it may be desirable to use more complicated and specialized preprocessing techniques. Preprocessing requirements are discussed at length in Ref. 6, but of course there are no techniques which are "optimum" except for specific parent distributions.

In summary, if the means and standard deviations of the raw measurement variables Z_{ji} over the set of observations to be used for establishing the regression surface are

denoted by $\bar{Z}_j$ and s_j, respectively, the usual normalizing necessary may be expressed by

$$X_{ji} = (Z_{ji} - \bar{Z}_j)/s_j \tag{3.7}$$

3.3 A FIRST-ORDER CORRECTION

The regression $E[Y|\underset{\sim}{X} = \underset{\sim}{x}]$ represents that function $h(\underset{\sim}{x})$ which minimizes the mean-squared error, $E[Y - h(\underset{\sim}{x})]^2$. However, even for large sample size, this objective is not realized by either $\hat{Y}(\underset{\sim}{x})$ or $Y_\ell^*(\underset{\sim}{x})$ because of systematic distortion of the estimated density which results when the smoothing parameter σ is greater than zero. In the case $n \to \infty$, the nature of this distortion is known since it is routine to show that

$$E[\hat{f}(\underset{\sim}{x})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\underset{\sim}{X}) g(\underset{\sim}{x} - \underset{\sim}{X})\, d\underset{\sim}{X} = f(\underset{\sim}{x}) * g(\underset{\sim}{x}) \tag{3.8}$$

where * indicates convolution,

$$g(\underset{\sim}{x}) \equiv (2\pi\sigma^2)^{-p/2} \exp\left(-\underset{\sim}{x}'\underset{\sim}{x}/2\sigma^2\right)$$

and $f(\underset{\sim}{x})$ is the probability density function of the distribution from which the sample is drawn.

If, for example, f is the normal distribution with mean μ and covariance Φ, as $n \to \infty$, $\hat{f}(\underset{\sim}{x})$ converges to a normal distribution with mean μ but with a covariance matrix of $[\Phi + \sigma^2 I]$, where I is the identity matrix. Since a covariance of $[\sigma^2 I]$ represents a distribution in which the variates are completely uncorrelated, addition of $\sigma^2 I$ to an arbitrary covariance Φ increases the variance terms with no effect on the covariance terms. This has the effect of biasing the estimated density in the direction of lower intercorrelations. Since the intercorrelations between the predicted

and predictor variables for the estimated density are characteristically less than these same intercorrelations for the parent density, the predictions for Y_i are characteristically closer to the mean than they should be. This effect has been noted in experience with real data.

As a simple but extreme example, consider the case of Y and X both normally distributed with zero mean, unit variance, and correlation one. Applying Eq. (2.4), it can be seen that as $n \to \infty$

$$\hat{Y}(x) \to \frac{\int_{-\infty}^{\infty} X \exp\left(-\frac{X^2}{2}\right) \exp\left[-\frac{(x - X)^2}{2\sigma^2}\right] dX}{\int_{-\infty}^{\infty} \exp\left(-\frac{X^2}{2}\right) \exp\left[-\frac{(x - X)^2}{2\sigma^2}\right] dX} = \frac{x}{\sigma^2 + 1} \qquad (3.9)$$

In this example, $E[Y|X = x] = x$ whereas $\hat{Y}(x)$ is only proportional to x and, as predicted, biased toward the mean. Similarly, it can be shown that a purely deterministic second-order component, $Y = AX^2$, is attenuated by the estimator $\hat{Y}(x)$ to yield

$$\hat{Y}(x) = A\left[\frac{x^2}{(\sigma^2 + 1)^2} + \frac{\sigma^2}{\sigma^2 + 1}\right]$$

As $\sigma \to 0$ both first- and second-order components are obtained without error, but with finite σ the error can be appreciable. However, the first-order scaling effect and the constant bias can be completely compensated. Once Eq. (2.4) or (3.5) is used to find a nonlinear relationship between $\underset{\sim}{X}$ and Y, the relationship between the resulting scalar $\hat{Y}$ and Y should be essentially linear. The best linear correction of $\hat{Y}$ (in the least-squares sense) is, of course, obtained through simple linear regression of Y on $\hat{Y}$.

Thus, a corrected estimate of Y could be obtained by

$$\hat{\hat{Y}}(\underset{\sim}{x}) = \alpha_o + \alpha_1 \hat{Y}(\underset{\sim}{x}) \tag{3.10}$$

where

$$\alpha_o = \left\{ \left[\sum Y_i \right] \left[\sum \left(\hat{Y}(\underset{\sim}{X}_i) \right)^2 \right] - \left[\sum \hat{Y}(\underset{\sim}{X}_i) \right] \left[\sum Y_i \hat{Y}(X_i) \right] \right\} \div D$$

$$\alpha_1 = \left\{ n \left[\sum Y_i \hat{Y}(\underset{\sim}{X}_i) \right] - \left[\sum Y_i \right] \left[\sum \hat{Y}(\underset{\sim}{X}_i) \right] \right\} \div D$$

$$D = n \sum [\hat{Y}(\underset{\sim}{X}_i)]^2 - \left[\sum \hat{Y}(\underset{\sim}{X}_i) \right]^2$$

and the summations run from $i = 1$ to $i = n$.

Section 4
COMPARISON WITH CONVENTIONAL TECHNIQUES

Nonlinear regression involves either a priori specification of the form of the regression equation with subsequent statistical determination of some undetermined constants, or statistical determination of the constants in a general regression equation – usually of polynomial form. The advantages and disadvantages of both approaches are well known. I will now point out some of the differences which distinguish the technique described in this paper.

The first approach requires that the form of the regressic i equation be known a priori or guessed. The advantage of this approach is that it usually reduces the problem to estimation of a relatively small number of undetermined constants, and that the values of these constants when found may provide some insight to the investigator. The disadvantage is that the regression is constrained to yield a "best fit" for the specified form of equation. If the specified form of equation is a poor guess and not actually appropriate to the data base to which it is applied, this constraint can be serious. Classical polynomial regression is usually limited to polynomials in one independent variable because polynomials involving multiple variates often have too large a number of free constants to be determined using a fixed number n of observations $(\underset{\sim}{X}_i, Y_i)$. A classical polynomial regression surface may fit the n-observed points very closely, but unless n is much larger than the number of coefficients in the polynomial, there is no assurance that the error $(Y^* - Y)$ for a new point $(\underset{\sim}{X}, Y)$ taken randomly from the distribution $f(\underset{\sim}{x}, y)$ will be small. On the other hand, with the regression Eq. (2.4) it is possible to let σ be small which allows high order curves if they are necessary to fit the data, but even in the limit as $\sigma \to 0$ Eq. (2.4) does not go wild but merely estimates $\hat{Y}(\underset{\sim}{X})$ as being the same as the Y_i associated with the $\underset{\sim}{X}_i$ which is closest in Euclidean distance to $\underset{\sim}{X}$. Cover (Ref. 5) points out that, for a wide range of probability distributions, the large-sample risk associated with estimation by this

"nearest-neighbor" rule is equal to only twice the Bayes risk (for squared error loss functions). For any $\sigma > 0$ there is a smooth interpolation between the observed points (as distinct from the discontinuous change of $\hat{Y}$ from one value to another at points equidistant from the observed points when $\sigma = 0$).

Since Eqs. (2.4) and (3.5) are mathematically identical when the polynomials are not truncated, the above statements are equally applicable to the polynomial regression equation of the form Eq. (3.5). The Eqs. (3.3) and (3.6) were derived in such a way that hundreds or thousands of terms can be introduced into the polynomial regression equation without overfitting the data even if the number of observed points is less than the number of coefficients. It is important to note that actual identity of the polynomial equation and Eq. (2.4) occurs when all possible terms are included in the polynomials. The significant point is that the polynomial approximation tends to fit the estimated regression surface given by Eq. (2.4), not the actual data. Equation (2.4), in turn, employs a density estimator which involves smoothing of the data – thereby minimizing the effects of randomness in sampling on the resulting regression surfaces. Thus, the number of terms used in the polynomials is limited only to minimize computation; there is no need to further limit this number because of any danger of overfitting the data. As a practical matter, the computed polynomials can usually be truncated to a low order with little degradation in the correlation between $\hat{Y}(\underset{\sim}{X}_i)$ and Y_i. If the dimensionality p is limited to 3 or 4, the number of terms can be held to a number which is easily manageable.*

A general formula for the computation time involved is not available but, as an example, one problem which has been run many times with different data has $p = 3$ and maximum order of terms in the polynomials $= 4$. The average time required to compute the 70 coefficients using 288 observations of $(X_{1i}, X_{2i}, X_{3i}, Y_i)$ and to evaluate $Y^*(\underset{\sim}{x})$ for each of the 288 observations was 690 milliseconds on the Univac 1108 computer.

---

*The total number of terms in a polynomial truncated to include terms up to the r order is given by Sebestyen (Ref. 7) to be $\binom{p+r}{p}$.

One additional feature of Eq. (2.4) is that $\hat{Y}(\underset{\sim}{x})$ is always bounded by the maximum and minimum values of the observed Y_i's. In contrast, the classical polynomial regression estimate goes to either ∞ or $-\infty$ as x goes to $\pm\infty$. Surprisingly, the polynomial-ratio regression estimate $Y^*(\underset{\sim}{x})$ is also bounded if $P_\ell(\underset{\sim}{x})$ and $Q_f(\underset{\sim}{x})$ are truncated to include only corresponding pairs of coefficients and enough terms are retained so that the contribution of each observation $\underset{\sim}{X}_i$ to $P_\ell(\underset{\sim}{x})$ is positive in the range of $\underset{\sim}{x}$ of interest. The question of bounds on $Y^*(\underset{\sim}{x})$ will be treated in more detail in the Appendix.

Section 5
INDEPENDENT VARIABLE NOT RANDOM

It has been pointed out that although the motivation used in arriving at Eq. (2.4) was based on both $\underset{\sim}{X}$ and Y being random variables, the concept of using for $\hat{Y}(\underset{\sim}{x})$ a weighted average of the Y_i's with the weight for each Y_i being some monotonically decreasing function of $|\underset{\sim}{X}_i - \underset{\sim}{x}|$ is attractive intuitively even if values of $\underset{\sim}{X}_i$ are specified in the design of an experiment. In this case, only the values Y_i of Y corresponding to the fixed values of $\underset{\sim}{X}_i$ would be measured observations of a random variable. If for the monotonically decreasing weighting function we choose

$$\exp\left[-|\underset{\sim}{X}_i - \underset{\sim}{x}|^2/2\sigma^2\right]$$

then the regression equation is given by Eq. (2.4). Since the derivation of the polynomial equivalent, Eq. (3.5), was not dependent on the assumption that $\underset{\sim}{X}$ is random, the computational advantages of the polynomial form can be utilized even when $\underset{\sim}{X}$ is not random, but instead values of $\underset{\sim}{X}$ are specified in the design of an experiment.

Section 6
REFERENCES

1. E. Parzen, "On Estimation of a Probability Density Function and Mode," Ann. Math. Statist., Vol. 33, 1962, pp. 1065–1076

2. T. Cacoullos, "Estimation of a Multivariate Density," Ann. Inst. Statist. Math., Vol. 18, No. 2, 1966, Tokyo, pp. 179–189

3. D. F. Specht, "Series Estimation of a Probability Density Function," 1968 (submitted for publication)

4. D. F. Specht, "Generation of Polynomial Discriminant Functions for Pattern Recognition," IEEE TRANS. on Elec. Comp., Vol. EC-16, 1967, pp. 308–319

5. T. M. Cover, Estimation by the Nearest-Neighbor Rule, SU-SEL-66-090 (TR 7002-1) Stanford Electronics Labs., Stanford, Calif., Sep 1966

6. D. F. Specht, Generation of Polynomial Discriminant Functions for Pattern Recognition, Ph.D. dissertation, Stanford University (also available as SU-SEL-66-029, Stanford Electronics Labs., Stanford, Calif., May 1966, and as Defense Documentation Center Report AD 487 537)

7. G. S. Sebestyen, Decision-Making Processes in Pattern Recognition, New York, Macmillan, 1962

Appendix

BOUNDS ON $Y_\ell^*(\underset{\sim}{x})$ FROM EQUATION (3.5)

From Eqs. (3.5) and (3.2)

$$Y_\ell^*(\underset{\sim}{x}) = \frac{\sum_{J \le \ell} b_{j_1 \cdots j_p} x_1^{j_1} \cdots x_p^{j_p}}{\sum_{J \le \ell} a_{j_1 \cdots j_p} x_1^{j_1} \cdots x_p^{j_p}} \tag{A.1}$$

Decomposing the coefficients into elements due to each observation

$$a_{j_1 \cdots j_p} = \sum_{i=1}^{n} \alpha_{j_1 \cdots j_p i}$$

$$\alpha_{j_1 \cdots j_p i} = \left[n\sigma^{2J} j_1! \cdots j_p! \right]^{-1} X_{1i}^{j_1} \cdots X_{pi}^{j_p} \exp\left(- \underset{\sim}{X}_i' \underset{\sim}{X}_i / 2\sigma^2 \right)$$

$$b_{j_1 \cdots j_p} = \sum_{i=1}^{n} \beta_{j_1 \cdots j_p i}$$

$$\beta_{j_1 \cdots j_p i} = \left[n\sigma^{2J} j_1! \cdots i_p! \right]^{-1} Y_i X_{1i}^{j_1} \cdots X_{pi}^{j_p} \exp\left(- \underset{\sim}{X}_i' \underset{\sim}{X}_i / 2\sigma^2 \right) = Y_i \alpha_{j_1 \cdots j_p i}$$

Then

$$Y^*_\infty(\underset{\sim}{x}) = \frac{\sum_{i=1}^{n}\left[\beta_{0\ldots0i} + \beta_{10\ldots0i}x_1 + \ldots + \beta_{j_1\ldots j_p i}x_1^{j_1}\ldots x_p^{j_p} + \ldots\right]}{\sum_{i=1}^{n}\left[\alpha_{0\ldots0i} + \alpha_{10\ldots0i}x_1 + \ldots + \alpha_{j_1\ldots j_p i}x_1^{j_1}\ldots x_p^{j_p} + \ldots\right]}$$

Let the expression contained in the first set of brackets be represented by $N_i^\infty(\underset{\sim}{x})$ and the expression contained in the second set of brackets be represented by $L_i^\infty(\underset{\sim}{x})$. If $N_i^\infty(\underset{\sim}{x})$ and $L_i^\infty(\underset{\sim}{x})$ are truncated to contain any arbitrary finite subsets of the original terms but contain only corresponding pairs of terms, then

$$N_i(\underset{\sim}{x}) = Y_i L_i(\underset{\sim}{x})$$

where $N_i(\underset{\sim}{x})$ and $L_i(\underset{\sim}{x})$ represent finite truncations of $N_i^\infty(\underset{\sim}{x})$ and $L_i^\infty(\underset{\sim}{x})$, respectively, and

$$Y^*(\underset{\sim}{x}) = \frac{\sum_{i=1}^{n} L_i(\underset{\sim}{x})Y_i}{\sum_{i=1}^{n} L_i(\underset{\sim}{x})} \qquad (A.2)$$

Thus, $Y^*(\underset{\sim}{x})$ is always equivalent to a weighted average of the Y_i's. The weightings are, of course, dependent on the value of $\underset{\sim}{x}$ at which $Y^*(\underset{\sim}{x})$ is to be evaluated.

When the weights $L_i(\underset{\sim}{x})$ are all nonnegative, the weighted average $Y^*(\underset{\sim}{x})$ is bounded by the maximum and minimum values of Y_i observed in the sample.

Since

$$L_i^{\infty}(\underset{\sim}{x}) = \frac{1}{n} \exp\left(-\frac{\underset{\sim}{X}_i'\underset{\sim}{X}_i}{2\sigma^2}\right)\left[1 + \frac{1}{\sigma^2}\underset{\sim}{X}_i'\underset{\sim}{x} + \frac{1}{2!\sigma^4}(\underset{\sim}{X}_i'\underset{\sim}{x})^2 + \ldots + \frac{1}{J!\sigma^{2J}}(\underset{\sim}{X}_i'\underset{\sim}{x})^J + \ldots\right]$$

$$= \frac{1}{n} \exp\left(\frac{\underset{\sim}{X}_i'\underset{\sim}{x} - \frac{1}{2}\underset{\sim}{X}_i'\underset{\sim}{X}_i}{\sigma^2}\right) \qquad \text{(A.3)}$$

$L_i(\underset{\sim}{x})$ is always positive in the untruncated case. Letting $z \equiv \underset{\sim}{X}_i'\underset{\sim}{x}/\sigma^2$

$$L_i(\underset{\sim}{x}) = \frac{1}{n} \exp\left(-\underset{\sim}{X}_i'\underset{\sim}{X}_i/2\sigma^2\right)\left[1 + z + \frac{z^2}{2!} + \ldots + \frac{z^{\ell}}{\ell!}\right] \qquad \text{(A.4)}$$

Since $\exp\left(-\underset{\sim}{X}_i'\underset{\sim}{X}_i/2\sigma^2\right) \geq 0$ for any $\underset{\sim}{X}_i$, $L_i(\underset{\sim}{x}) \geq 0$ if

$$S_{\ell}(z) \equiv 1 + z + \frac{z^2}{2!} + \ldots + \frac{z^{\ell}}{\ell!} \geq 0$$

But

$$\frac{dS_{\ell}(z)}{dz} = S_{\ell-1}(z)$$

and

$$S_{\ell}(z) = \frac{z^{\ell}}{\ell!} - \frac{d}{dz} S_{\ell}(z)$$

Since the minimum value of $S_{\ell}(z)$ occurs when $\frac{d}{dz} S_{\ell}(z) = 0$,

$$\min S_{\ell}(z) = \frac{z^{\ell}}{\ell!} \quad \text{for some value of } z \quad -\infty \leq z \leq \infty$$

Thus

$$\min S_\ell(z) \geq 0 \qquad \ell \text{ even} \tag{A.5}$$

and therefore $L_i(\underline{x}) \geq 0$ for all z when $P_\ell(\underline{x})$ and $Q_\ell(\underline{x})$ are truncated to include all terms up to the ℓ order and ℓ is even.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Lockheed Palo Alto Research Laboratory Lockheed Missiles & Space Company Palo Alto, California	Unclassified
	2b. GROUP N/A

3. REPORT TITLE

A PRACTICAL TECHNIQUE FOR ESTIMATING GENERAL REGRESSION SURFACES

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Lockheed Independent Research Report

5. AUTHOR(S) (First name, middle initial, last name)

Donald F. Specht

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1968	22 text pages	7

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N/A b. PROJECT NO.	6-79-68-6
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Distribution of the document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	N/A

13. ABSTRACT

Let $\underset{\sim}{X}$ be a p component random vector variable with transpose $\underset{\sim}{X}' = [X_1, \ldots, X_j, \ldots, X_p]$ and Y be a one-dimensional random variable with a joint continuous distribution of density $f(\underset{\sim}{x}, y)$. The regression of Y given $\underset{\sim}{X} = \underset{\sim}{x}$ is

$$E[Y|\underset{\sim}{X} = \underset{\sim}{x}] = \int_{-\infty}^{\infty} y f(\underset{\sim}{x}, y)\, dy \div \int_{-\infty}^{\infty} f(\underset{\sim}{x}, y)\, dy \qquad (1)$$

Using consistent nonparametric estimators of the densities, $E[Y|X = x]$ can be approximated by a ratio of two general polynomials where the coefficients of the polynomials are computed as a function of the observed sample. The technique described has the following advantages compared with classical polynomial regression techniques utilizing a single polynomial: 1. The shape of the regression surface can be made as complex as necessary to closely approximate Eq. (1) or as simple as desired, by proper choice of a single parameter σ (called the smoothing parameter). In spite of this flexibility, the estimated value is bounded by the minimum and maximum of the observations Y_i when the polynomials are truncated at an even order. 2. Because of the smoothing properties inherent in the density estimator, the number of coefficients used in the polynomials can approach or even exceed the number of observations in the sample with no danger of the regression surface overfitting the data when σ is suitably chosen. 3. A simple method of determining the coefficients has been derived. The calculation for the coefficient of a particular term amounts to little more than averaging the corresponding product of variables over the set of observations available. 4. The computational and storage requirements increase only linearly with the number of coefficients used. The computational advantages of the technique can be utilized also when values of $\underset{\sim}{X}$ are specified in the design of an experiment and Y alone is a random variable.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Regression surfaces						
Nonlinear regression						
Polynomial-ratio						
Smoothing						
Interpolation						
Computational techniques						
Density estimators						